SEPARATING SIGNAL FROM NOISE: A SELF-DISTILLATION APPROACH FOR AMORTIZED HETEROGENEOUS CRYO-EM RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cryogenic electron microscopy (cryo-EM) is a powerful method for visualizing heterogeneous biomolecular structures at high resolution, but the growing scale of cryo-EM datasets creates substantial computational challenges for 3D reconstruction. Autoencoder-based heterogeneous reconstruction methods offer a potential solution through amortized inference: once trained, an encoder could infer the conformations of new particle images with a single forward pass. In practice, however, encoders of existing methods overfit to the extremely noisy images used for training and fail to generalize to unseen particles. Here, we introduce **Cryo**-**N**o-**O**verfit (CryoNOO), a training paradigm for learning noise-invariant representations that generalize to unseen cryo-EM images. CryoNOO leverages the reconstruction model itself as a denoiser, reconstructing and reprojecting training particles to generate denoised views that are augmented with independent noise realizations for self-supervised learning. Across synthetic and experimental datasets, we show that CryoNOO enables large-scale amortized reconstruction from a subset of particles, generalizes to conformations absent during training, and supports accurate conformation inference on unseen particles.

1 INTRODUCTION

Single-particle cryo-electron microscopy (cryo-EM) has revolutionized structural biology by enabling the determination of macromolecular structures at near-atomic resolution (Nakane et al., 2018; Yip et al., 2020). Beyond determining individual structures, cryo-EM can directly capture structural heterogeneity within experimentally observed molecular populations, providing access to conformational ensembles underlying biological processes such as allosteric regulation, enzyme catalysis, and ligand binding. The growing use and throughput of cryo-EM have led to a rapid increase in the scale of cryo-EM datasets over the past two decades (Iudin et al., 2023; Subramaniam et al., 2026).

Recovering heterogeneous structures from cryo-EM data presents a challenging inverse problem: each particle image is an extremely noisy 2D projection of a single 3D conformation in an unknown orientation. A growing class of deep-learning methods tackles heterogeneous reconstruction using autoencoder-like architectures (Donnat et al., 2022), in which an encoder maps each particle image to a low-dimensional representation of its conformation and a decoder maps this representation to a 3D structure whose projections are compared with the observed images.

A key potential advantage of these architectures is *amortized inference*. Once trained, an encoder could infer the conformational representation of a previously unseen particle with a single forward pass, allowing a reconstruction model trained on a subset of particles to be applied efficiently to much larger datasets. In practice, however, encoders in existing heterogeneous reconstruction methods overfit to the particle images used for training and fail to generalize to unseen images (Edelberg & Lederman, 2023). Because cryo-EM images have extremely low signal-to-noise ratios, an encoder can learn image-specific noise rather than the underlying structural signal: a model may accurately reconstruct structures represented in its training set while producing poor or uninformative latent representations for new images of the same structures. This failure limits the ability of existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

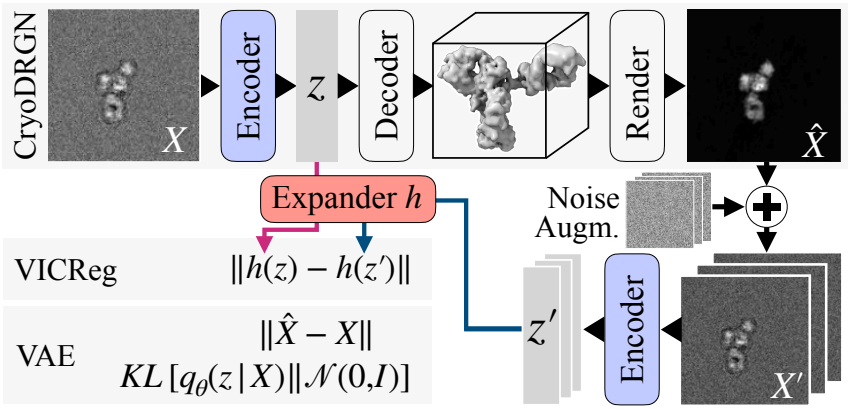

Figure 1: **Method overview.** CryoNOO (Cryo-NO-Overfit) learns noise-invariant representations through self-distillation in autoencoder-based cryo-EM reconstruction. A training image X is passed through the CryoDRGN VAE to produce a latent representation z and a denoised projection $\hat{X}$ rendered from the reconstructed 3D structure. We augment $\hat{X}$ with independent realizations of additive Gaussian noise to generate views X', which are passed through the same encoder to obtain latent representations z'. An expander MLP h maps the latents to a representation space where VICReg encourages invariance across noise realizations while maintaining variance and reducing covariance across dimensions. The model is trained end-to-end with the CryoDRGN VAE and VICReg objectives.

heterogeneous reconstruction methods to realize the computational and representational benefits of amortized inference.

In this work, we introduce Cryo-NO-Overfit (CryoNOO), a training paradigm for autoencoder-based heterogeneous cryo-EM reconstruction that learns representations that generalize to unseen particle images. Our key insight is to leverage the reconstruction model itself to separate structural signal from image-specific noise and generate additional views for self-supervision. Given a noisy training image, we reconstruct its 3D structure and reproject it to synthesize a denoised 2D view, which can then be augmented with independent noise realizations. We encourage the encoder to map the original and noise-augmented views to consistent latent representations, thereby promoting invariance to image-specific noise. To prevent representational collapse, we use Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2021), which combines an invariance objective with variance and covariance regularization.

We demonstrate that learning noise-invariant representations enables several capabilities beyond standard heterogeneous reconstruction. CryoNOO enables large-scale amortized reconstruction by training on only a subset of a dataset and inferring representations for the remaining particles; its latent space generalizes to conformations absent during training; and its learned representations support accurate amortized conformation inference on unseen particles. Together, these results demonstrate that noise-invariant representation learning can overcome encoder overfitting and enable generalizable amortized inference for heterogeneous cryo-EM reconstruction.

2 Related Works

Amortized Cryo-EM Methods. Neural network-based cryo-EM reconstruction methods commonly employ VAE architectures to model continuous structural heterogeneity via amortized inference. CryoDRGN (Zhong et al., 2020; 2021), for example, learns a function $\hat{V} : \mathbb{R}^{3+n} \to \mathbb{R}$ that maps a low-dimensional latent space to a 3D electron scattering potential in the Fourier domain. While cryoDRGN has the potential for amortized inference of conformations, in practice it fails to generalize to unseen views and tends to memorize the training images (Edelberg & Lederman, 2023; Levy et al., 2025). In the *ab initio* setting, CryoAI (Levy et al., 2022a), CryoFire (Levy et al., 2022b), and CryoSPIN (Shekarforoush et al., 2024) all investigate amortized inference of poses by jointly predicting the pose. However, unlike our work these works do not study held-out general-

ization. Amortized model have also been investigated for template matching: CryoSBI (Dingeldein et al., 2025) uses simulation-based inference to learn per-image distributions over templates.

Recent neural models (Punjani & Fleet, 2023; Levy et al., 2024; 2025) circumvent possible memorization by dispensing with the encoder altogether, instead using an autodecoder architecture. Unlike autoencoders, where an encoder maps 2D views to latent codes, autodecoders do not have an encoder and instead give each view a unique learnable latent code, stored in a lookup table.

Self-Supervised Learning and Self-Distillation. In self-supervised learning from images, we are interested in learning an encoder that produces latent representations useful for downstream tasks using only unlabeled images. This is usually accomplished by encoding augmentations of the same image and enforcing them to be similar in latent space. The core challenge is to avoid trivial solutions, such as predicting a constant representation independent of the encoded image. This can be accomplished in different ways. In this paper, we rely on VICReg (Bardes et al., 2021) which avoids representational collapse by enforcing three principles: encouraging consistency across views, maintaining diversity across samples, and reducing redundancy across feature dimensions. Also related to our method are methods of self-distillation, where a "student" encoder is distilled from a "teacher" whose weights are however an exponential moving average of the student's weights (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021). Our method can be seen as a form of self-distillation, where we are distilling knowledge already inherent in the CryoDRGN model into a more powerful encoder, though we rely on VICReg to prevent representational collapse. VICReg has no notion of generating pseudo-groundtruth with a 3D reconstruction model, and to the best of our knowledge, VICReg has never been used in combination with a differentiable forward model or renderer.

3 METHOD

3.1 PRELIMINARIES

Cryo-EM Image Formation Model. Cryo-electron microscopy (cryo-EM) aims to reconstruct a 3D molecular structure $V : \mathbb{R}^3 \to \mathbb{R}$ from noisy 2D projection images $X_1, \ldots, X_N$, each corresponding to a particle in an unknown pose defined by a rotation $R \in SO(3)$ and an in-plane translation $t \in \mathbb{R}^2$. The image formation model can be written as $X_i = C_i\, P_{\phi_i} V + \eta_i$, where C_i models the contrast transfer function (CTF), P_{ϕ_i} denotes the projection operator for pose $\phi_i = (R_i, t_i)$, and $\eta_i \sim \mathcal{N}(0, \sigma^2)$ is additive noise. In practice, reconstruction is typically performed in Fourier space using the Fourier Slice Theorem (FST) (Bracewell, 1956), which shows that each 2D image corresponds to a central slice of the 3D Fourier transform of V.

CryoDRGN as an implicit teacher. CryoDRGN models structural heterogeneity with a VAE where the encoder $q_\xi(\mathbf{z}|X)$ with parameters ξ maps each particle image to a latent vector $\mathbf{z}$ representing structural variability. This latent representation is passed to a conditional implicit neural representation (INR) decoder to reconstruct a 3D structure, which is then projected into the 2D image domain via the cryo-EM image formation model (Sec. 3.1) and compared to the input image using a standard VAE loss. Although trained with a variational objective, CryoDRGN effectively *denoises*: it reconstructs a clean 3D structure from noisy 2D inputs. This observation enables us to view the model itself as a *teacher* that can supervise later iterations of its own encoder – a form of *self-distillation*.

VICReg. To prevent representation collapse in self-supervised learning, VICReg (Bardes et al., 2021) introduced a loss composed of three terms: invariance, variance and covariance regularization. These operate on the representation space produced by an extra expander network h to enforce non-triviality and disentanglement across embedding dimensions.

The *variance* term encourages each embedding dimension to exhibit sufficient variation across the batch using a hinge function, while the *covariance* term penalizes correlations between dimensions:

$$v(z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0,\, \gamma - \sqrt{\mathrm{Var}_{\text{batch}}(h(z_j)) + \epsilon}\right), c(z) = \frac{1}{d}\sum_{i\neq j}\left[\frac{1}{n-1}\, h(Z)^\top h(Z)\right]_{i,j}^2 . \quad (1)$$

where $\gamma = 1$ is the target standard deviation, ϵ is a small constant for numerical stability, and $Z \in \mathbb{R}^{n\times d}$ is the batch of centered embeddings. The *invariance loss* enforces the consistency

between embeddings $h(z)$ and $h(z')$ of different augmentations of the same input:

$$s(z, z') = \frac{1}{n} \sum_{i=1}^{n} \|h(z_i) - h(z'_i)\|_2^2. \tag{2}$$

The total VICReg loss combines these terms:

$$\mathcal{L}_{\text{VICReg}}(z, z') = \lambda_{\text{inv}}\, s(z, z') + \lambda_{\text{var}}\big(v(z) + v(z')\big) + \lambda_{\text{cov}}\big(c(z) + c(z')\big). \tag{3}$$

where $\lambda_{\text{inv}} = \lambda_{\text{var}} = \lambda_{\text{cov}} = 1$ by default. VICReg has primarily been used for image classification, object detection, and retrieval. To the best of our knowledge, this is the first instance of self-supervised learning techniques applied to cryo-EM reconstruction to enable generalization.

3.2 CryoNOO

In Cryo-No-Overfit (CryoNOO), we propose a new self-supervised learning objective that enables heterogeneous cryo-EM reconstruction methods to reconstruct images not in the training set. While our method can in principle work with any encoder-decoder model, we use the variational autoencoder (VAE) architecture of CryoDRGN. Our method can be viewed as a simple extension that enables the encoder to separate signal from noise.

Definitions. We define a CryoDRGN model as an encoder $q_\xi(\mathbf{z}|X)$ that takes an image X and produces a latent $\mathbf{z}$ along with a decoder $p_\theta(\mathbf{z}, k)$ that takes a latent $\mathbf{z}$ and coordinate $k = (k_x, k_y, k_z)$ in Fourier space and produces the Fourier transform of the electron scattering at k. ξ and θ are the weights of the encoder and decoder, respectively. In cryo-EM reconstruction, we are given a noisy input image $X = T(I)$, where $T \sim \mathcal{T}$ is a transformation that applies additive Gaussian noise $\mathcal{N}(0, \sigma)$ in real space, and I is the underlying noiseless image. Both T and I are unknown.

Denoising and Self-Distillation. Our key insight is that CryoDRGN inherently acts as a denoiser, as it produces clean 3D reconstructions from noisy input images. We leverage this denoising property to use CryoDRGN as its own teacher, performing self-distillation in the form of generating noise-augmented views for self-supervised learning. In order to ensure that CryoDRGN is a strong denoiser, we first pre-train the model for N epochs with the usual VAE objective. Then, in subsequent epochs, we perform denoising as follows: given a noisy input image $X = T_0(I), T_0 \sim \mathcal{T}$, we compute its latent representation $\mathbf{z}_0 = q_\xi(T_0(I))$, reconstruct a denoised 3D volume using decoder p_θ, and render a 2D image $\hat{X}$ using the cryo-EM image model (see Eqn. 8 and Fig. 1). Since $\hat{X}$ is in Fourier space, we first apply the inverse Hartley transform to convert $\hat{x}$ to a real-space image $\hat{I}$, an estimate of the clean image I. We do not stop gradients through $\hat{X}$. We then add Gaussian noise $\epsilon \sim N(0, f_j\sigma)$, where noise factors $f_j, j = \{1, \ldots, J\}$ are hyperparameters, and apply the forward Hartley transform to obtain augmented views $X' = T_1(\hat{I}), \ldots, T_J(\hat{I})$. The noise level σ is estimated from the training set (Appendix A).

Self-supervised learning for noise-invariant representations. With multiple augmented views $T(I) = T_0(I), \ldots, T_J(I)$, we perform self-supervised learning to encourage robust and noise-invariant representations. Each augmented image is passed through the encoder q_ξ to obtain latent vectors $\mathbf{z}_1, \ldots, \mathbf{z}_J$. To enforce invariance across these noise-augmented views while preserving meaningful structural variation, we apply the VICReg (Bardes et al., 2021) loss, chosen for both its empirical effectiveness and architectural simplicity. Unlike contrastive approaches, VICReg does not require negative samples or separate encoders for different views. In addition to enforcing invariance, VICReg promotes decorrelation across embedding dimensions and enforces sufficient per-dimension variance, which helps prevent representation collapse and supports disentanglement of conformational variability.

To use VICReg, we simply add a projection head h (expander), implemented as a multi-layer perceptron (MLP), which maps each latent vector $\mathbf{z}_j$ to a higher-dimensional embedding space. The VICReg objective (see Section 3.1) is jointly optimized with the original VAE loss during training. The overall training objective is:

$$\mathcal{L}(X, \mathbf{z}_0, \ldots, \mathbf{z}_J) = \mathcal{L}_{\text{VAE}}(X, \mathbf{z}_0) + \lambda_{\text{VICReg}} \sum_{j=1}^{J} \mathcal{L}_{\text{VICReg}}(\mathbf{z}_0, \mathbf{z}_j) \tag{4}$$

Table 1: **Large-scale amortized reconstruction.** Comparison between CryoDRGN and CryoNOO evaluated on all 1M images. CryoDRGN-1M serves as a reference for the full-data regime. Best results in **bold**; ↑ higher is better, ↓ lower is better; CD = Chamfer distance.

	Train	Train	Test	IgG-1D			Tomotwin-100		
Method	images	time	poses?	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)
CryoDRGN	1M	10×	Yes	0.397	1.043	0.681	0.389	1.380	0.709
CryoNOO	100K	1.3×	No	**0.376**	**1.286**	**0.641**	**0.317**	**4.749**	**0.596**
CryoDRGN	100K	1×	No	0.335	5.793	0.498	0.279	6.553	0.542

where $\lambda_{\mathrm{VICReg}}$ is a hyperparameter that controls the strength of the VICReg loss. We use 0.3 for this $\lambda_{\mathrm{VICReg}}$. As in VICReg, the expander h is only used for training and is thrown away after training, so after training we are left with a CryoDRGN model and inference proceeds identically to that of CryoDRGN. An overview of the full procedure is shown in Figure 1.

4 Results

In our experiments, we apply CryoNOO to cryoDRGN, a state-of-the-art method for heterogeneous cryo-EM reconstruction. We design three experiments to evaluate the benefits of learning noise-invariant representations. First, we show that CryoNOO enables large-scale amortized reconstruction, reducing computational cost by training on only a subset of particles and inferring representations for the remainder with a single encoder pass (Sec. 4.1). Second, we test whether these representations generalize to conformations entirely unseen during training (Sec. 4.2). Finally, we evaluate whether the learned representations support amortized conformation inference on unseen particles (Sec. 4.3). Section 4.4 ablates the components of CryoNOO to identify the sources of these improvements.

4.1 Large-scale amortized reconstruction

For modern cryo-EM datasets containing millions to tens of millions of particles, amortization offers a substantially cheaper alternative to processing every particle. In this setting, consensus refinement and model training are performed on only a subset of particles, with the trained encoder performing inference on the remaining particles. Because poses enter only through the forward model used in the training loss, inference requires no pose estimation: each unseen particle is mapped to its latent code (and hence its reconstructed volume) with a single encoder pass.

Datasets and metrics We generate large-scale versions of the `IgG-1D` and `Tomotwin-100` datasets from CryoBench (Jeon et al., 2024) with 1M images following their original generation protocols. The `IgG-1D` dataset simulates a 1D circular motion of a Fab domain of the IgG antibody, and the `Tomotwin-100` dataset simulates a heterogeneous mixture of 100 distinct proteins curated from the TomoTwin dataset (Rice et al., 2023). All methods are trained with ground-truth poses for the training subset, isolating the effect of amortization from pose-estimation error.

Reconstructions are evaluated against ground truth maps, using the following metrics: area under the Fourier Shell Correlation (FSC) curve ($\mathrm{FSC}_{\mathrm{AUC}}$), real-space Chamfer Distance (CD), and real-space volumetric intersection-over-union (vIoU) (Gu et al., 2026). More details on $\mathrm{FSC}_{\mathrm{AUC}}$ can be found in the Appendix E.

Results In Table 1, we compare CryoNOO and cryoDRGN models trained on a 100K subset of the particles and amortized to the remaining 900K particles. A cryoDRGN model trained on all 1M particles is included as a reference for the full-data regime. By training on only 10% of the particles and amortizing inference over the remainder, CryoNOO-100K requires only $1.3\times$ the training time of cryoDRGN-100K, compared with $10\times$ for cryoDRGN-1M, corresponding to a $8.7\times$ reduction in training time relative to processing the full dataset with cryoDRGN.

Despite this substantially lower computational cost, CryoNOO-100K achieves considerably better reconstruction quality over the full 1M-particle dataset than cryoDRGN-100K. Compared with

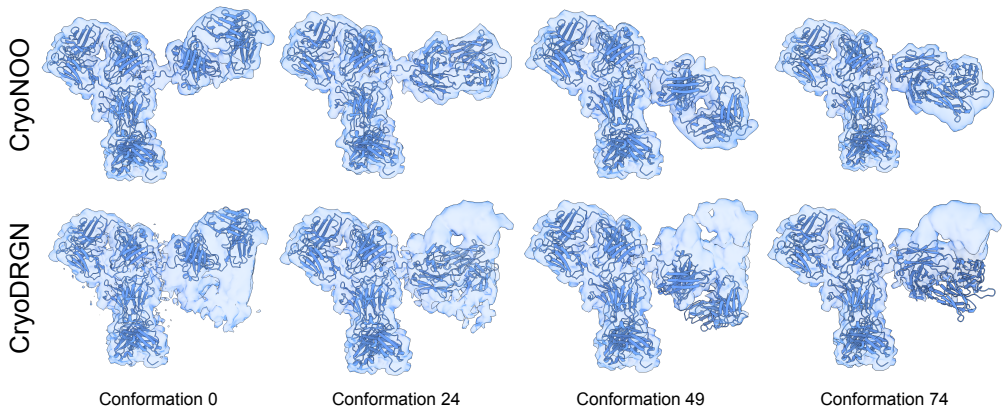

Figure 2: **CryoNOO generalizes to unseen particles.** Test-set reconstructions from CryoNOO-100K and cryoDRGN-100K for four conformations of `IgG-1D`-1M, overlaid with the corresponding ground-truth atomic structures (blue).

cryoDRGN-1M, CryoNOO-100K closes much of the performance gap on `IgG-1D`, the conformational heterogeneity setting, while a larger gap remains on `Tomotwin-100`, the compositional heterogeneity setting.

In Figure 2, we visualize four randomly sampled test reconstructions from `IgG-1D` for CryoNOO-100K and cryoDRGN-100K, overlaid on the corresponding ground-truth atomic structures. CryoNOO's reconstructed maps closely match the ground-truth structures, whereas cryoDRGN's reconstructions show poorer agreement and substantial artifacts, including small disconnected densities near the Fab domain and, frequently, two fused Fab domains instead of one. Further experiments, including validation of this pipeline on experimental data, are provided in Appendix C.1.

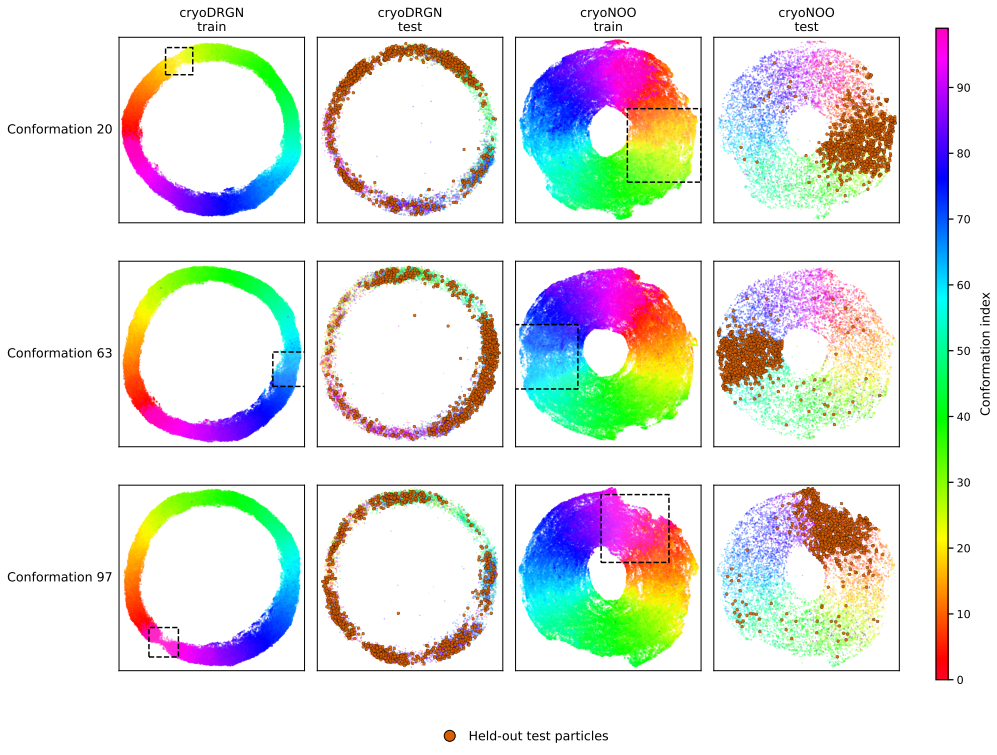

Figure 3: **Latent space generalization to unseen conformations.** Latent representations of training particles and particles from three held-out conformations (indices 20, 63, and 97) for cryoDRGN (left) and CryoNOO (right). Boxes indicate the expected latent-space locations of the held-out conformations.

Table 2: **Amortized conformation inference.** ↑ higher is better, ↓ lower is better. Top-5 (window) counts a hit if the true conformation falls within ±2 of the point estimate; top-5 (rank) if it is among the 5 highest-ranked conformations. cMAE = circular mean absolute error.

Method	Top-1 Acc. (↑)	Top-5 (rank) (↑)	Top-5 (window) (↑)	cMAE (↓)
CryoDRGN (Zhong et al., 2020)	12.44	56.68	47.70	3.612
CryoSBI (Dingeldein et al., 2025)	21.36	75.78	67.00	2.422
CryoNOO	**22.34**	**81.00**	**72.84**	**1.619**

4.2 Latent space generalization to unseen conformations

For amortized reconstruction to recover the full conformational landscape of a dataset, the encoder must generalize not only to unseen particle images, but also to conformations that may be absent from the subset used for training. In particular, a semantically meaningful latent space should place particles from a previously unseen conformation according to their relationship to conformations observed during training, rather than mapping them onto existing states.

Experimental setup To test this form of generalization, we use the `IgG-1D` dataset (Jeon et al., 2024), in which conformations vary continuously along a periodic one-dimensional reaction coordinate. We remove particles corresponding to one conformation from the training set and train CryoNOO on the remaining conformations. We then perform inference on a test set containing particles from the held-out conformation. Because the neighboring conformations remain present during training, the held-out state has a well-defined location between them along the underlying conformational coordinate.

Results We evaluate generalization by examining where particles from the held-out conformation are embedded relative to the conformations observed during training. In Figure 3, cryoDRGN fails to consistently localize particles from the held-out conformation, instead distributing their latent codes across regions occupied by observed conformations. In contrast, CryoNOO places most held-out particles in the latent-space region between the appropriate neighboring conformations, recovering their expected position along the conformational coordinate despite never observing this state during training. We observe similar behavior when different conformations are held out, indicating that this generalization is not specific to a particular region of the conformational landscape.

4.3 Amortized conformation inference

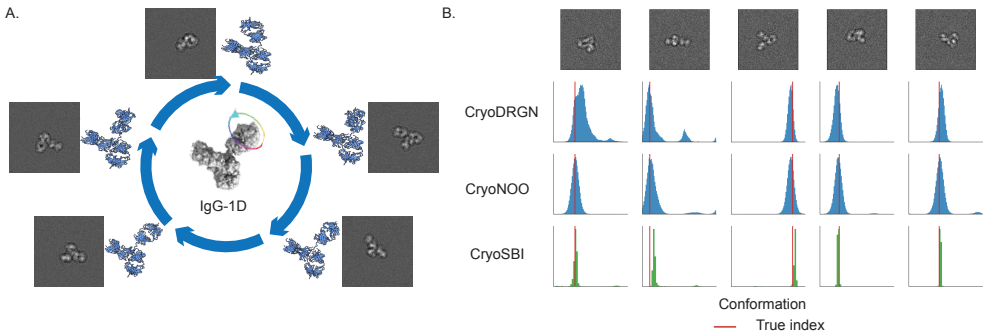

Figure 4: **Amortized conformation inference. A.** `IgG-1D` conformational landscape with representative particle images. **B.** Posterior distributions over conformation for five held-out test particle images for each method.

Beyond reconstructing unseen particles, a generalizable encoder should produce representations that support downstream inference about molecular conformation. We therefore ask whether the latent representations learned by CryoNOO can be used to infer the conformational state of previously unseen particles. Given labeled conformations for the training particles, we fit a lightweight probe on the frozen latent representations and use it to infer conformations for held-out particles. Because the latent representation itself is inferred by the encoder, conformation inference is amortized: each new particle requires only a forward pass through the encoder followed by the probe.

We compare against cryoSBI (Dingeldein et al., 2025), which formulates this problem as amortized template matching. Given a set of candidate conformations parameterized by a reaction coordinate θ, cryoSBI uses a cryo-EM simulator to generate particle images from atomic models and directly trains a conditional density estimator on simulated $(\mathbf{x}, \theta)$ pairs to approximate the posterior $q(\theta \mid \mathbf{x})$. In contrast, cryoDRGN and CryoNOO learn their representations through heterogeneous reconstruction without access to conformation labels. We use labels only after training to probe whether the learned representations support conformation inference.

Since cryoDRGN and CryoNOO are VAEs, their encoders provide a natural representation for constructing a surrogate posterior over conformational states. The encoder maps a particle image $\mathbf{x}$ to a latent representation $\mathbf{z}$ via $q_\xi(\mathbf{z} \mid \mathbf{x})$. Suppose we are given a discrete set of conformations parameterized by $\theta_1, \ldots, \theta_K$. Assuming that the latent distribution for each conformation is an equal-covariance isotropic Gaussian centered at μ_i, we define $q(\mathbf{z} \mid \theta_i) = \mathcal{N}(\mathbf{z} \mid \mu_i, \sigma^2 I)$, where $\mathbf{z}$ is the latent code produced by the VAE encoder and the centroids μ_i are estimated from labeled training particles. Assuming a uniform prior over conformations, the corresponding posterior is

$$p(\theta_i \mid \mathbf{x}) := p(\theta_i \mid \mathbf{z}) = \frac{\exp\left(-\|\mathbf{z} - \mu_i\|^2 / 2\sigma^2\right)}{\sum_j \exp(-\|\mathbf{z} - \mu_j\|^2 / 2\sigma^2)}. \tag{5}$$

Under these assumptions, assigning a particle to its nearest centroid is Bayes-optimal.

For continuous or ordered conformations, we instead construct a non-parametric posterior from the k nearest training-set neighbors of $\mathbf{z}$ in latent space. Specifically, we approximate the posterior over conformations by the empirical distribution of their labels and obtain a point prediction from the mean of the k-NN labels, with k selected by 5-fold cross-validation. This allows both models to perform amortized conformation inference while also providing an estimate of uncertainty.

Dataset and metrics We evaluate conformation inference on a simulated version of `IgG-1D` generated using the cryoSBI simulator, with CTF simulation following CryoBench (Jeon et al., 2024). We compare cryoDRGN, cryoSBI, and CryoNOO on their ability to infer the conformational state of held-out particles. We evaluate discrete prediction using top-1 and top-5 accuracy and continuous prediction using mean absolute error (MAE). Because the `IgG-1D` conformational coordinate is periodic (Figure 4A), we compute error using the circular distance $d(\hat{y}, y) = \min\left(|\hat{y} - y| \bmod 100,\ 100 - (|\hat{y} - y| \bmod 100)\right)$ and report the resulting circular MAE (cMAE). We report two notions of top-5 accuracy: standard top-5 rank accuracy and a continuous-window metric that considers a prediction correct when the true conformation lies within ± 2 states of the point estimate. The continuous k-NN posterior is only used to compute the cMAE.

Results Table 2 reports conformation-inference performance on a held-out test set. Despite learning its representation without conformation labels, CryoNOO achieves higher top-1 and top-5 accuracy and lower cMAE than cryoSBI, which is trained directly to infer conformation. Both methods substantially outperform cryoDRGN. Because cryoDRGN and CryoNOO use the same VAE architecture and downstream probe, their comparison directly demonstrates that the noise-invariant training objective produces representations that better encode conformational state.

Figure 4B visualizes the inferred posterior distributions for five randomly selected test particles. CryoNOO and cryoSBI generally produce localized posteriors near the true conformation, whereas cryoDRGN produces broader, less localized distributions. In these examples, cryoSBI tends to produce narrower posteriors, but several place substantial probability away from the true conformation. Together with the quantitative results, these examples illustrate that CryoNOO's unsupervised latent representations support accurate amortized conformation inference on unseen particles.

4.4 Ablations

In Table 3, we ablate the main hyperparameters and design choices of CryoNOO on the original `IgG-1D` dataset (Jeon et al., 2024). Performance is relatively robust to several hyperparameter choices; for example, expander dimensions of 1024 and 2048 yield comparable results. Varying the noise magnitude by $0.5\times$ and $2\times$ also results in only modest degradation, indicating that CryoNOO does not require precise estimation of the experimental noise level.

Table 3: **Ablation study on IgG-1D.** All variants evaluated using test per-image $\mathrm{FSC}_{\mathrm{AUC}}$ (higher is better). **Ours** denotes the chosen configuration; the best-scoring variant is underlined.

Variant	$\mathrm{FSC}_{\mathrm{AUC}}$	Variant	$\mathrm{FSC}_{\mathrm{AUC}}$
Ours	**0.347**	Expander dim = 32	0.298
w/o CryoDRGN warmup	0.338	Expander dim = 128	0.340
+ warmup (N=5)	0.344	Expander dim = 1024 (default)	0.347
+ warmup (N=40)	0.337	Expander dim = 2048	0.348
Single noise (J=1)	0.344	Expander dim = 4096	0.345
2x noise	0.342	No Expander	0.296
0.5x noise	0.340	Only Invariance	0.325
$\lambda = 1.0$	0.339		
$\lambda = 0.1$	0.345		
SimSiam (Chen & He, 2021)	0.332	CryoDRGN (Zhong et al., 2020)	0.325
CryoDRGN + Noise Aug	0.328		

We next examine whether the gains from CryoNOO can be reproduced by simpler forms of noise invariance or alternative self-supervised objectives. Replacing VICReg with SimSiam (Chen & He, 2021), another self-supervised objective that does not require a separate momentum encoder, reduces performance. Simpler approaches perform substantially worse: directly augmenting particle images with additional noise does not provide the same benefit, while enforcing only agreement between the latent representations of two independent noise realizations ("Only Invariance") is insufficient. Together, these ablations indicate that the gains from CryoNOO arise not simply from exposure to additional noise or an invariance constraint alone, but from the full self-supervision objective used to learn noise-invariant representations.

5 Discussion

We introduce CryoNOO, to our knowledge the first method to explicitly address encoder overfitting in amortized heterogeneous cryo-EM reconstruction. Leveraging the inherent denoising capability of VAE-based reconstruction methods such as cryoDRGN, CryoNOO uses self-supervision on independent noise realizations to learn noise-invariant latent representations. We show that these representations generalize to particle images unseen during training, enabling accurate amortized reconstruction where a standard cryoDRGN encoder struggles. Beyond reconstruction, CryoNOO's latent representations generalize to conformations absent from the training set and support accurate amortized conformation inference on unseen particles.

This generalization enables a different approach to scaling heterogeneous reconstruction to increasingly large cryo-EM datasets. Rather than training on every particle, CryoNOO can be trained on a subset of the dataset and then applied to the remaining particles through inexpensive encoder inference. In our million-particle experiments, training CryoNOO on only 10% of the particles substantially reduces training cost while retaining accurate reconstruction of the full dataset. More broadly, these results suggest that learning generalizable encoders can decouple the computational cost of model training from the total number of particles to which the model is ultimately applied.

Although we demonstrate CryoNOO with cryoDRGN's conformation encoder, the proposed self-supervision is not specific to a particular encoder architecture or latent variable. The same approach could in principle be applied to amortized pose inference or other cryo-EM inference tasks in which image noise encourages an encoder to memorize individual observations. More broadly, generalizable amortized inference may become increasingly important as cryo-EM datasets continue to grow and novel imaging paradigms enable data collection at larger scales or introduce new inference settings. We therefore view noise-invariant representation learning as a general strategy for improving amortized inference in cryo-EM, enabling models trained on a subset of a dataset to support reconstruction and downstream inference across substantially larger collections of particles.

AI use statement In this work, we used generative AI tools to implement and edit parts of the codebase and to help data analysis through the suggestion of metrics. We have not used generative AI tools for generating synthetic datasets, processing datasets, proposing hypotheses, or to design experiments, and mathematical claims and proofs are not applicable to this work. Additionally, we

used generative AI tools to summarize and find existing literature, source information, and edit and review the paper. We have reviewed all AI-assisted work. LLM-generated code was reviewed by the authors. Summarized and sourced information was verified with the original source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement For further details on the datasets used in this paper, see Appendices A and C.1 and Section 4.3. For details on the method, see Section 3. For additional details on noise estimation, see Appendix A. Code will be released upon acceptance of the paper.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2): 198–217, 1956.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Lars Dingeldein, David Silva-Sánchez, Luke Evans, Edoardo D'Imprima, Nikolaus Grigorieff, Roberto Covino, and Pilar Cossio. Amortized template matching of molecular conformations from cryoelectron microscopy images using simulation-based inference. *Proceedings of the National Academy of Sciences*, 122(23):e2420158122, 2025.

Claire Donnat, Axel Levy, Frederic Poitevin, Ellen D Zhong, and Nina Miolane. Deep generative modeling for volume reconstruction in cryo-electron microscopy. *Journal of structural biology*, 214(4):107920, 2022.

Daniel G Edelberg and Roy R Lederman. Using vaes to learn latent variables: observations on applications in cryo-em. *ArXiv*, pp. arXiv–2303, 2023.

J Ryan Feathers, Robert Heeter, Geoffrey Woollard, Sonya M Hanson, Pilar Cossio, Joel Greer, Tom Burnley, and Ellen Zhong. The cahra challenge: A community-wide assessment of cryo-em heterogeneous reconstruction algorithms. *bioRxiv*, pp. 2026–09, 2026.

Ryan Feathers, Erica K Richael, Kayla A Simanek, J Christopher Fromme, and Jon E Paczkowski. Structure of the rhlr-pqse complex from pseudomonas aeruginosa reveals a new mechanism of quorum sensing transcriptional regulation. *Pseudomonas aeruginosa*, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Jeffrey Gu, Minkyu Jeon, Ambri Ma, Serena Yeung-Levy, and Ellen D Zhong. Cryohype: Reconstructing a thousand cryo-em structures with transformer-based hypernetworks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35280–35290, 2026.

Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023.

Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M Hanson, Pilar Cossio, and Ellen D Zhong. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em. *arXiv preprint arXiv:2408.05526*, 2024.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. 2022a. URL https://arxiv.org/abs/2203.08138.

Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022b.

Axel Levy, Rishwanth Raghu, David Shustin, Adele Peng, Huan Li, Oliver Clarke, Gordon Wetzstein, and Ellen Zhong. Mixture of neural fields for heterogeneous reconstruction in cryo-em. *Advances in Neural Information Processing Systems*, 37:56988–57017, 2024.

Axel Levy, Rishwanth Raghu, J Ryan Feathers, Michal Grzadkowski, Frederic Poitevin, Francesca Vallese, Oliver B Clarke, Gordon Wetzstein, and Ellen D Zhong. Cryodrgn-ai: Neural ab initio reconstruction of challenging cryo-em and cryo-et datasets. *bioRxiv*, pp. 2024–05, 2025.

Takanori Nakane, Dari Kimanius, Erik Lindahl, and Sjors HW Scheres. Characterisation of molecular motions in cryo-em single-particle data by multi-body refinement in relion. *elife*, 7:e36861, 2018.

Ali Punjani and David J Fleet. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870, 2023.

Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. Tomotwin: generalized 3d localization of macromolecules in cryo-electron tomograms with structural data mining. *Nature methods*, 20(6):871–880, 2023.

Shayan Shekarforoush, David Lindell, Marcus A Brubaker, and David J Fleet. Cryospin: improving ab-initio cryo-em reconstruction with semi-amortized pose inference. *Advances in Neural Information Processing Systems*, 37:55785–55809, 2024.

Sriram Subramaniam, Werner Kühlbrandt, and Richard Henderson. Cryo-em: the revolution continues. *IUCrJ*, 13(4), 2026.

Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial intelligence and statistics*, pp. 384–391. PMLR, 2009.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *International conference on artificial neural networks*, pp. 485–491. Springer, 2001.

Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-em. *Nature*, 587(7832):157–161, 2020.

Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-em images. In *International Conference on Learning Representations*, 2020.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

A Implementation Details

Optimization. We train all models in Pytorch using 1 NVIDIA H100 GPU. For the synthetic datasets, we follow CryoDRGN (Zhong et al., 2020; 2021) using the same training schedule and optimization parameters. We use Adam (Kingma, 2014) optimizer with learning rate set to 0.0001 without weight decay. Moreover, we applied learning rate scheduling using `torch.optim.lr_scheduler.ReduceLROnPlateau`, which reduces the learning rate when the total loss stops decreasing. Specifically, we used the scheduler in `min` mode with a reduction factor of 0.5 and a patience of 3 epochs. This scheduling strategy was only applied to CryoNOO, as no improvement was observed when applied to CryoDRGN (Zhong et al., 2020).

Noise estimation for experimental datasets. Following the cryo-EM image formation model (Eqn. 8), we assume that the noise in cryo-EM images is additive mean zero Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ in real space. Thus, to create noise-augmented images that resemble the data distribution from denoised images, it suffices to estimate σ of the Gaussian additive noise.

Our strategy for estimating σ consists of masking the low-frequency signal and estimating σ as the standard deviation of the high-frequency noise. We derive our noise estimation as follows: first, we note that almost all of the power of a cryo-EM image is concentrated at the low frequencies in the power spectrum. Thus, the high frequencies of the power spectrum are dominated by the power spectrum of Gaussian noise, which has a constant power spectrum σ^2. To compute σ, we calculate the power spectrum of the image and average the power at all frequencies k bigger than some threshold k_{thresh}. We can then calculate σ using the Plancherel theorem for the 2D discrete Fourier transform as $\sigma = \sqrt{\frac{P}{D^2}}$. The complete algorithm is given in Algorithm 1. The `radial_average` function takes a 2D array and averages the values of the ray over concentric rings around the center of the 2D array.

Algorithm 1 Noise estimation from cryo-EM images

Require: Cryo-EM image $I_n \in \mathbb{R}^{D \times D}, n = 1, \ldots, N$, frequency threshold k_{thresh}

$\sigma\text{s} \leftarrow []$
for $j \leftarrow 1, \ldots, N$ **do**
  $\hat{I}_j \leftarrow \text{FFT}(I_j)$
  $PS \leftarrow |\hat{I}_j|^2$
  $r \leftarrow \texttt{radial_average}(PS)$
  $P \leftarrow \frac{1}{D - k_{\text{thresh}}} \sum_{k=k_{\text{thresh}}}^{D} r[k]$
  $\sigma\text{s} \leftarrow \sigma\text{s} \cup \{\sqrt{P/D^2}\}$
end for
return $\sigma = \frac{1}{N} \sum_{i=1}^{N} \sigma\text{s}[i]$

Noise factors (f_j). We multiply each noise factor f_j with the base standard deviation σ to generate Gaussian noise, i.e., $\epsilon \sim N(0, f_j\sigma)$. For synthetic datasets, we can reproduce the exact noise level by simulating images using the image formation model described in Jeon et al. (2024), starting from the `before_noise` images. Since the original noise level (σ) was estimated from 256×256 images and we downsampled them to 128×128, the effective noise level increases by a factor of 2. Therefore, applying a *noise factor* of 2 restores the original noise level. The detailed noise factors (f_j) and noise levels (σ), including the original σ for synthetic datasets and the estimated σ for experimental datasets, are shown in Table 4.

Dataset generation for large-scale amortized reconstruction Starting from $256 \times 256 \times 256$ density maps provided by CryoBench (Jeon et al., 2024), we generate 2D projections by uniformly sampling poses $R \in SO(3)$ and in-plane translations $t \in [20, 20]^2$ pixels. The contrast transfer function (CTF) is modeled with an accelerating voltage of 300 kV, a spherical aberration at 2.7 mm, and an amplitude contrast of 0.1. Defocus values are randomly sampled without replacement from EMPIAR-11247 (Feathers et al., 2022). Gaussian noise is added to achieve a signal-to-noise (SNR) ratio of 0.1. We simulate 10,000 images per conformation, resulting in a total of 1M images. Finally, the images are downsampled to $D = 128$ using Fourier cropping.

Experimentally-collected datasets. For experimentally-collected datasets, we estimate the noise using Algorithm 1 as well as by computing the statistics of noise in the background (i.e.. edges) of the images. These two estimates often agree to at least the first decimal place, but where they differ we use the second estimate. We also observe that the estimated noise σ often appears to be an underestimate to the true noise σ, with significantly better performance at higher σs. We hypothesize that this is because in experimental data, high-frequency noise is attenuated due to detector MTF and dose-weighted motion correction. We recommend using a noise factor of 10 as a rule of thumb, and use this factor on all of our experiments on real data. For the CAHRA datasets, all estimated noise levels were similar, consistent with their shared acquisition, so we used ADH's estimated noise of $\sigma = 2.509$ for all CAHRA datasets.

Table 4: **Noise standard deviation (sigma) for each dataset.**

Dataset	**Original / Estimated σ**	**Noise factors f_j**
IgG-1D (SNR 0.1)	1.6979	1, 2, 3
IgG-1D (SNR 0.01)	5.3662	1, 2, 3
IgG-RL (SNR 0.01)	5.3662	1, 2, 3
Ribosembly (SNR 0.01)	11.2579	1, 2, 3
Tomotwin-100 (SNR 0.01)	6.2319	1, 2, 3
Spike-MD (SNR 0.01)	327.1266	2
All CAHRA datasets	2.509	10

Architecture. We adopt the standard CryoDRGN (Zhong et al., 2020) architecture, consisting of a VAE with fully connected encoder and decoder networks. Both the encoder and decoder are 3-layer residual MLPs with 1024 hidden dimensions, and the latent variable z is 8-dimensional. For the expander h, we use a 3-layer MLP with 1024 hidden dimensions per layer. BatchNorm and ReLU activations are applied after each layer, except the last layer.

Masking. In the original CryoDRGN setting, a circular spatial mask is applied during both decoding and loss computation for computational efficiency and improved performance, while the encoder processes the full image. In contrast, in our setting, the mask is applied only during loss computation, as we retain the full image size in order to add noise after the inverse Fourier transform.

B Computational costs

Table 5 compares CryoDRGN (Zhong et al., 2020) and our proposed CryoNOO in terms of memory usage and training time. We report GPU memory, CPU memory, and wall-clock time along with other relevant factors. The performance gap between using a single noise factor and three noise factors in CryoNOO is relatively small, so we include results for both settings. We find that in the common resolution (box) 128 setting CryoNOO takes approximately 1.3x more wall-clock time and 1.4x more GPU memory, though these costs are resolution-dependent and increase with the resolution. CPU memory cost consists almost entirely of loading the training dataset into memory and is therefore the same between both methods.

C Additional Experiments

C.1 Amortized Reconstruction

We evaluate our method on CryoBench (Jeon et al., 2024), a suite of synthetic cryo-EM reconstruction datasets covering a range of different types of heterogeneity, and the CAHRA Heterogeneity Challenge (Feathers et al., 2026). To test generalization outside of the training data, we partition the data into training and test sets with a 90/10 split.

Table 5: **Computational cost by dataset.** Peak GPU memory, peak host RSS, and wall-clock compute per epoch, with the per-epoch ratio relative to CryoDRGN (1×) on the same dataset. Box (D) gives the box size in voxels with the padded dimension used for training in parentheses.

Dataset	Box (D)	Method	Noise	Peak GPU mem	Peak host RSS	Compute/ epoch	Rel. time
IgG-1D	128 (129)	CryoDRGN	–	8.1 GB	11.7 GB	0:57	1×
		CryoNOO	3 (1,2,3)	11.4 GB	11.7 GB	1:17	1.35×
IgG-RL	128 (129)	CryoDRGN	–	8.1 GB	11.7 GB	0:56	1×
		CryoNOO	1	11.4 GB	11.7 GB	1:23	1.48×
		CryoNOO	3 (1,2,3)	11.4 GB	11.7 GB	1:14	1.32×
Ribosembly	128 (129)	CryoDRGN	–	8.1 GB	11.7 GB	3:10	1×
		CryoNOO	3 (1,2,3)	11.4 GB	11.7 GB	4:14	1.34×
Tomotwin-100	128 (129)	CryoDRGN	–	8.1 GB	11.7 GB	0:56	1×
		CryoNOO	3 (1,2,3)	11.4 GB	11.7 GB	1:16	1.36×
Spike-MD	256 (257)	CryoDRGN	–	35.4 GB	24.2 GB	4:05	1×
		CryoNOO	1 (2)	45.1 GB	24.3 GB	5:09	1.26×
		CryoNOO	3 (1,2,3)	54.4 GB	24.3 GB	5:38	1.38×

C.1.1 DATASETS

CryoBench CryoBench consists of five synthetic datasets: `IgG-1D` and `IgG-RL`, which simulate a 1D circular motion and random orientations of the Fab domain of human IgG, `Ribosembly`, which simulates ribosome assembly, `Tomotwin-100`, which simulates a complex mixture of 100 distinct proteins, and `Spike-MD`, which simulates a realistic motion of the SARS-CoV2 spike protein generated via molecular dynamics (MD) simulation.

CAHRA Heterogeneity Challenge We use a subset of proteins from the CAHRA heterogeneity challenge (Feathers et al., 2026), which is an experimentally-collected dataset of proteins with similar structure. The individual proteins are combined to form a complex mixture that is challenging for cryo-EM reconstruction methods. From the full CAHRA dataset, we use ADH, Aldolase, PKM2, and a mixture of the three proteins to assess our method. The CAHRA dataset provides disjoint high quality and low quality particle subsets. High quality and low quality particles were generated by using the first and last 30 frames of each movie, respectively, so lower quality particles have higher accumulated dose and radiation damage. For our experiments, poses were obtained by the CAHRA dataset's consensus refinement of the HQ particles only.

We assess amortization with two experiments: HQ → HQ, where we train and test solely on the high quality particles of each dataset, and HQ → LQ, where we train on the high quality particles of each dataset and perform amortized inference on the low quality particles of each dataset.

Results In Table 6, we find that cryoDRGN and CryoNOO are comparable when evaluated on particles seen during training, with cryoDRGN perhaps having a small numerical advantage due to its ability to overfit to the training set. On test particles, CryoNOO significantly outperforms cryoDRGN on all reconstruction metrics, showing the greater amortization of our combined self-distillation and self-supervised learning approach. Numerically, CryoNOO shows bigger improvements on datasets with compositional heterogeneity (Ribosembly, Tomotwin-100). However, this effect is difficult to disentangle from encoder misidentification being more costly in the compositional heterogeneity setting.

In Table 7, we find similar gains in amortization on the HQ → HQ experiment, showing that CryoNOO's ability to amortize extends to experimentally-collected data. In the HQ → LQ experiment, we validate that our proposed pipeline of performing consensus refinement on a subset of the particles, training a reconstruction model on that subset, and performing inference on the rest of the particles can work in practice on experimentally-collected data.

Table 6: **Quantitative performance on CryoBench.** Comparison between CryoDRGN and CryoNOO on train and test splits across CryoBench. ↑ higher is better, ↓ lower is better; CD = Chamfer distance. Best test-split result per dataset in **bold**.

		Train			Test		
Dataset	**Method**	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)
IgG-1D	CryoDRGN	0.353	1.279	0.646	0.325	4.570	0.519
	CryoNOO	0.356	1.434	0.630	**0.347**	**1.950**	**0.594**
IgG-RL	CryoDRGN	0.315	2.999	0.599	0.309	6.401	0.514
	CryoNOO	0.321	3.296	0.575	**0.317**	**5.206**	**0.521**
Ribosembly	CryoDRGN	0.407	0.833	0.782	0.346	4.944	0.621
	CryoNOO	0.402	0.875	0.776	**0.356**	**4.033**	**0.657**
Tomotwin-100	CryoDRGN	0.327	1.626	0.694	0.165	13.60	0.369
	CryoNOO	0.326	1.640	0.691	**0.239**	**8.310**	**0.508**
Spike-MD	CryoDRGN	0.335	3.346	0.470	0.305	2.676	0.490
	CryoNOO	0.331	3.307	0.470	**0.307**	**2.136**	**0.505**

Table 7: **Quantitative performance on CAHRA, high-quality (HQ) and HQ→LQ transfer.** Mean per-image FSC-AUC, Chamfer distance (CD), and volumetric IoU (vIoU) over the reconstructed volumes, on individual and combined datasets. **(a)** Models trained and evaluated on high-quality (HQ) data. **(b)** The same models evaluated on low-quality (LQ) data, i.e. HQ→LQ transfer. ↑ higher is better, ↓ lower is better.

		Train			Test		
Dataset	Model	AUC (↑)	CD (↓)	vIoU (↑)	AUC (↑)	CD (↓)	vIoU (↑)
(a) High-quality (HQ → HQ)							
ADH	CryoDRGN	0.338	1.067	0.436	0.356	0.929	0.453
	CryoNOO	**0.433**	**0.578**	**0.659**	**0.432**	**0.582**	**0.655**
Aldolase	CryoDRGN	0.297	1.348	0.378	0.339	**0.973**	0.428
	CryoNOO	**0.344**	**1.314**	**0.487**	**0.382**	0.982	**0.560**
PKM2	CryoDRGN	0.344	1.544	0.502	0.377	1.095	0.565
	CryoNOO	**0.387**	**0.803**	**0.576**	**0.402**	**0.682**	**0.606**
Combined	CryoDRGN	0.400	0.751	0.607	0.335	1.132	0.530
	CryoNOO	**0.420**	**0.599**	**0.617**	**0.356**	**1.120**	**0.538**
(b) High-quality (HQ)→ low-quality (LQ) transfer							
ADH	CryoDRGN	0.333	1.143	0.426	0.370	2.055	0.477
	CryoNOO	**0.409**	**0.722**	**0.593**	**0.421**	**0.618**	**0.620**
Aldolase	CryoDRGN	0.305	**1.238**	0.376	0.345	2.238	0.445
	CryoNOO	**0.347**	1.305	**0.490**	**0.377**	**0.992**	**0.549**
PKM2	CryoDRGN	0.345	1.550	0.503	0.389	0.812	0.596
	CryoNOO	**0.389**	**0.804**	**0.577**	**0.405**	**0.625**	**0.607**
Combined	CryoDRGN	0.400	0.726	0.609	0.348	**1.074**	0.546
	CryoNOO	**0.426**	**0.565**	**0.637**	**0.357**	1.095	**0.556**

C.1.2 Large-scale Amortization

In Table 8, we show the full train and test results for amortization. Table 1 combined the train and test metrics in order to compare performance between partial training with amortization versus full training. We find that although cryoDRGN-100K and CryoNOO-100K are comparable on train

Table 8: **Amortization.** Additional amortization metrics. ↑ higher is better, ↓ lower is better. CD = Chamfer distance. Parenthetical values in the lower block report medians.

	Train			Test		
Model	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)	$\mathrm{FSC}_{\mathrm{AUC}}$ (↑)	CD (↓)	vIoU (↑)
IgG-1D						
CryoDRGN 100k	0.385	1.108	0.665	0.329	6.313	0.479
CryoNOO 100k	0.384	1.120	0.663	**0.375**	**1.304**	**0.638**
Tomotwin-100						
CryoDRGN 100k	0.361	1.541	0.685	0.270	7.110 (2.599)	0.515 (0.611)
CryoNOO 100k	0.369	1.482	0.689	**0.311**	**5.112** (1.613)	**0.584** (0.668)

Table 9: **Latent consistency.** KLD, TV, JSD, and Hellinger distance depend on kernel density estimation; Wasserstein distance (WD) does not. Lower is better; best per column within each dataset in **bold**.

		KDE-based				Non-KDE
Dataset	**Method**	KLD	TV	JSD	Hell.	WD
IgG-1D	CryoDRGN	8.7917	0.7183	0.3733	0.6688	1.7896
	CryoNOO	**2.7179**	**0.5288**	**0.2091**	**0.4829**	**1.5085**
IgG-RL	CryoDRGN	14.4223	0.8796	0.5448	0.8405	1.9433
	CryoNOO	**4.4206**	**0.5460**	**0.2294**	**0.5086**	**1.6903**
Ribosembly	CryoDRGN	2.2327	**0.1787**	**0.0422**	**0.1714**	1.5359
	CryoNOO	**2.2074**	0.2560	0.0824	0.2367	**1.2979**
Tomotwin-100	CryoDRGN	9.0205	0.2425	0.0651	0.2241	2.2033
	CryoNOO	**4.7882**	**0.1675**	**0.0297**	**0.1561**	**1.6903**
Spike-MD	CryoDRGN	13.0702	0.8770	0.5425	0.8422	1.8361
	CryoNOO	**5.8449**	**0.6839**	**0.3386**	**0.6478**	**1.8084**

reconstruction metrics, we find significant gains in all test reconstruction metrics for CryoNOO-100K, mirroring our CryoBench results (Table 6).

C.2 Latent Space Generalization

Traditional cryo-EM reconstruction methods such as cryoDRGN are brittle, likely due to the high noise characteristic of cryo-EM images. Edelberg & Lederman (2023) finds that simple changes to a cryo-EM image can greatly shift its latent code, sometimes across clusters. Generalizable models should preserve latent space structure, with train and test latent spaces having the same organization.

Train-test per-conformation overlap We propose a measure of latent space structure preservation which assesses whether each part (i.e. conformation) of latent space is in a similar location with a similar distribution. For structures with K different states, we propose to measure the overlap in training and test latent distributions as the average of the difference between the distributions of each of the K states:

$$d_{\mathrm{latent}}(\mathbb{Z}_{\mathrm{train}}, \mathbb{Z}_{\mathrm{test}}) = \frac{1}{K}\sum_{i=1}^{K} d_{\mathrm{dist}}(\mathbb{Z}_{\mathrm{train,i}}, \mathbb{Z}_{\mathrm{test,i}}) \tag{6}$$

The latent distributions $\mathbb{Z}_{\{\mathrm{train,test}\},\mathrm{i}}$ are defined from the latent embeddings of each state i via a Gaussian kernel density estimation (KDE), and log-likelihoods of both train and test samples are evaluated. From these, we compute the measure of distributional difference d_{dist} with KL divergence, total variation distance, Jensen-Shannon divergence, Hellinger distance, and Wasserstein distance (L2). The first four metrics require KDE, while Wasserstein distance does not require forming a KDE and can be computed directly between the latent train and test point clouds. KDE

Table 10: **Bandwidths used in distribution shift experiments.**

Dataset	Method	Train Bandwidth	Test Bandwidth
IgG-1D	CryoDRGN	0.5	0.5
	CryoNOO	0.3	0.3
IgG-RL	CryoDRGN	0.7	0.5
	CryoNOO	0.7	0.7
Ribosembly	CryoDRGN	0.4	0.4
	CryoNOO	0.4	0.4
Tomotwin-100	CryoDRGN	0.3	0.7
	CryoNOO	0.3	0.5
Spike-MD	CryoDRGN	0.5	0.5
	CryoNOO	1.1	0.7

Table 11: **Latent quality.** Latent-space quality metrics for unseen test data on IgG-1D, Ribosembly, and Tomotwin-100. Continuous metrics apply to IgG-1D; cluster metrics apply to the distinct-cluster datasets. Higher is better; best per column in **bold**.

		Continuous		Cluster		
Dataset	**Method**	Trust.	Cont.	ARI	AMI	k-NN
IgG-1D	CryoDRGN	0.6308	0.6385	–	–	–
	CryoNOO	**0.8580**	**0.7911**	–	–	–
Ribosembly	CryoDRGN	–	–	0.1968	0.3607	38.72
	CryoNOO	–	–	**0.3121**	**0.4553**	**50.32**
Tomotwin-100	CryoDRGN	–	–	0.2097	0.4337	35.71
	CryoNOO	–	–	**0.5020**	**0.6714**	**67.30**

bandwidth is computed with cross-validation on a grid of bandwidths using likelihood as the metric (Table 10). For states with continuous heterogeneity (e.g. Spike-MD), we view these distributions as only having a single class ($K = 1$).

Table 9 shows that CryoNOO preserves latent space structure much better between the train and test distributions than cryoDRGN. This effect persists across metrics that require KDE as well as metrics that do not require KDE. For all CryoBench datasets besides Ribosembly, CryoNOO is better in all metrics. For Ribosembly, KDE and non-KDE metrics disagree. This shows that CryoNOO's latent spaces are more consistent between train and test particles, providing evidence that it reflects the underlying distribution rather than just the training sample.

Kernel width in KDE. In Table 9, we presented the alignment between the train and test latent distributions using k-fold cross-validation, where the bandwidth search space ranges from 0.1 to 10 in intervals of 0.2. The log-likelihood of the data is optimized during cross-validation. The selected bandwidths are reported in Table 10.

Measuring latent space quality In this section, we measure the quality of a latent space independent of any reference distributions. For distributions which have an ideal structure in latent space, such as a ring for IgG-1D, we propose to measure the quality of the latent space by measuring how well the latent space preserves the local structure of the ideal structure. The local preservation metrics we use are trustworthiness and continuity (Venna & Kaski, 2001; Van Der Maaten, 2009), which measure if neighbors in the latent space are neighbors in the ideal structure and if points that are neighbors in the ideal structure stay neighbors in the latent space, respectively. The formula for

trustworthiness:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i(k)} \big(r(i,j) - k\big), \qquad U_i(k) = N_k^Z(i) \setminus N_k^X(i) \tag{7}$$

where k is the number of neighbors, $r(i,j)$ is the rank of j in the list of nearest neighbors of i in the ideal structure X, and $U_i(k)$ is the set of k-nearest neighbors in latent space Z that is not in the set of k-nearest neighbors in the ideal structure X. Continuity is the same measure with the two spaces Z, X flipped. Trustworthiness and continuity range from $[0, 1]$ with higher indicating better local structure preservation. In Table 11, we find that CryoNOO better preserves IgG-1D's ring structure on unseen test data compared to CryoDRGN. For latent spaces that are comprised of many distinct clusters, such as Tomotwin-100 or Ribosembly, we use Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and k-NN classification accuracy to evaluate them.

In Table 11, we find that CryoNOO consistently generates better-quality latent spaces than cryoDRGN on unseen test data, regardless of conformational heterogeneity or compositional heterogeneity and dataset. In particular, CryoNOO also reports better quality on Ribosembly, despite mixed results on consistency. Combined with the previous result on latent space preservation, CryoNOO has both higher-quality and more consistent latent spaces and thus more generalizable latent spaces.

D CRYO-EM IMAGE FORMATION MODEL

The goal of cryo-electron microscopy (cryo-EM) reconstruction is to recover a 3D molecular structure $V : \mathbb{R}^3 \to \mathbb{R}$ from a collection of noisy 2D projection images $X_1, \ldots, X_N$. Each projection corresponds to an individual particle captured in an unknown pose, parameterized by a rotation $R \in SO(3)$ and an in-plane translation $t \in \mathbb{R}^2$. The image formation process for each observation X_i can be modeled in real space as:

$$X(r_x, r_y) = g * \left(\int_{\mathbb{R}} V(R^\top r)\, dr_z + t \right) + \epsilon, \tag{8}$$

where $r = (r_x, r_y, r_z)^\top$, $t = (t_x, t_y)$ is the translation vector, and ϵ denotes additive noise introduced during imaging. The function $g : \mathbb{R}^2 \to \mathbb{R}$ represents the microscope's point spread function (PSF), which is convolved with the 2D projection of the 3D volume. In practice, reconstruction is generally performed in Fourier space by leveraging the Fourier Slice Theorem (FST) (Bracewell, 1956), which simplifies the computation of the image formation model by replacing the projection integral with taking a slice through the 3D Fourier transform of the volume.

E PER-IMAGE FSC FOR HETEROGENEOUS RECONSTRUCTION

Following Jeon et al. (2024), we use per-image FSC to jointly evaluate reconstruction quality and structural heterogeneity. In cryo-EM, the Fourier Shell Correlation (FSC) curve is a standard tool for comparing two volumes by measuring their correlation across concentric shells in the Fourier domain. Since cryo-EM methods often reconstruct a 3D volume from a single 2D image, per-image FSC quantifies how well each individual reconstruction matches the ground truth conformation. Specifically, we compute the FSC between the reconstructed volume from a single image and its corresponding ground truth, then summarize performance using the area under the FSC curve ($\mathrm{FSC}_{\mathrm{AUC}}$). To capture the distribution of reconstruction quality across conformations, we sample one image per conformation as in Jeon et al. (2024). The maximum value of $\mathrm{FSC}_{\mathrm{AUC}}$ is 0.5.